AN EFFICIENT FRAMEWORK FOR PREDICTING AND FORECASTING OF INFECTIOUS DISEASES USING MACHINE LEARNING TECHNIQUES IN BIG DATA ENVIRONMENT

A SYNOPSIS

Submitted by

K. INDHUMATHI
(Reg. No. 9200017001)

in partial fulfillment for the award of the degree of

DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER APPLICATION

SCHOOL OF COMPUTING

KALASALINGAM ACADEMY OF RESEARCH AND EDUCATION

(DEEMED TO BE UNIVERSITY)

ANAND NAGAR, KRISHNANKOIL – 626 126

TAMIL NADU, INDIA

JULY 2023

1. INTRODUCTION:

Infectious diseases are diseases that are caused by pathogens (viruses, bacteria and parasites etc). The pathogen can spread infectious diseases to humans in several ways. The disease spread may occur from human to human. Otherwise, it happens through mosquitos, sewages, air pollution, water pollution and environmental condition. Environmental condition plays a vital role in spreading infectious diseases because the longevity, reproduction, dissemination and transmission of bacteria need appropriate climate and weather conditions [1]. Unexpected weather condition affects the entire resources of the environment [2]. So, the changes in climate or weather conditions may cause infectious diseases among humans [2].

The dengue virus is the cause of dengue fever, a tropical illness spread by mosquitoes. It is the Aedes aegypti mosquito that is responsible for the DENV virus, which is the cause of dengue fever. It is the most common viral infection spread by mosquitoes in the world today and a severe threat to global health. An estimated 3 billion people worldwide are at risk of disease, with 40 million symptom episodes and about 20,000 reported yearly fatalities. The incidence of dengue has increased dramatically in recent decades, with estimates of 40-50% of the world's population at risk for the disease in tropical, subtropical, and, most recently, more temperate areas.

Malaria is a fatal illness that primarily affects tropical nations. It is treatable as well as preventative. However, a case of uncomplicated Malaria can develop into a severe disease if quick diagnosis and suitable treatment are not provided. Malaria is carried through the bites of female Anopheles mosquitoes but is not contagious and cannot be passed from one person to another. The risk of Malaria affects almost half of the world's population. In 85 countries in 2021, an estimated 247 million people acquired Malaria. Around 619 000 people died from the illness in the same year.

Salmonella bacteria are the source of typhoid fever, often known as enteric fever. Salmonella is a movable Enterobacteriaceae that can cause a range of gastrointestinal illnesses. It can manifest in many ways, from an intense septic infection to small bouts of diarrhoea with low-grade fever. As of 2019, there are thought to be 9 million cases of Typhoid and 110 000 fatalities. Medications can treat typhoid fever, but the process becomes more challenging as different medicines become less effective.

A bacterial, viral, or fungal lung infection is known as Pneumonia. Pus or fluid may collect in your lungs due to Pneumonia, which inflames the lung tissue. Viral Pneumonia frequently goes away independently and is typically less severe than bacterial Pneumonia. Both lungs may be impacted by Pneumonia. Over 700,000 children under five die from Pneumonia yearly, or over 2,000 per day, making it the infectious disease that claims the lives of most kids. Over 200,000 births are included in this.

Every year the affected ratio of infectious diseases is very high. The symptoms of infectious disease may vary depending on the age group. So, the researcher faces difficulties in predicting which disease creates the highest impact among the several contagious diseases, and also identifying the symptom patterns depending on the age group of these infectious diseases is very difficult.

This thesis proposes new disease prediction and forecasting models to predict and forecast the spread of infectious diseases. This research developed an Integrated Random Forest XG-Boost model to find the highest occurrence of infectious diseases and then proposed a hybrid classification model PRFBXG-BOOST and CHAID to classify the risk factors of the infectious diseases based on the patient's age. Finally, this research developed an efficient time series forecasting model based on climate data to forecast the next 15 years of disease occurrence.

2. REVIEW LITERATURE:

2.1 Disease prediction using machine learning:

Sharma et al.[3] was used a bragging-based ensemble forecasting model to anticipate the number of disease incidences based on past occurrences. Log transformation and z-score transformations were used in the preprocessing. A specific disease dataset was used to demonstrate the applicability of this ensemble model. This model was tested for dengue, tuberculosis, chickenpox, and food poisoning, among other infections.

In Grampurohit & Sagarnal [4], a comparative analysis of machine learning classifiers' performance was presented based on disease prediction. Classifiers, including Naive Bayes, Decision Trees and Random Forest, were utilized to develop the disease prediction system. The symptoms from the obtained datasets for 41 different

diseases were used to evaluate this disease prediction algorithm.AI approaches were employed to forecast diseases using a feature selection method. Classifiers such as decision trees, SVM, random forest, and XG-Boost algorithms were utilized to classify diseases like Malaria, Jaundice, COVID, Pneumonia, Dengue and Typhoid, Dutta et al. [5].

2.2 Symptom prediction using machine learning:

Ayo et al. [6] identifies typhoid fever using ANOVA based on patient health records. It provides 97% accuracy for prediction. Thary & Azidan [7], the study identifies typhoid fever based on patient symptoms and lab blood tests. It used machine learning algorithms for predicting typhoid infections. The algorithms are weighted k-nearest neighbor and naïve bayes. Weighted KNN provides 98% accuracy, and naïve bayes provides 84% accuracy for typhoid prediction. The weighted knn performs well as compared to other algorithms. Mariki et al.[8] ,the researchers applied random forest for predicting malaria occurrence based on patient symptoms and population data.

Dataset has 2556 patient records with 36 features such as temperature, life span, tiredness, head pain, etc. This study used five machine learning algorithms: linear regression, k- nearest neighbour, random forest, support vector machine and decision tree. The random forest model produced 95% accuracy compared to machine learning models. Adamu & Singh [9] presents a predictive model for malaria incidence depending on the infected patient records. Climate elements and non-climatic elements are included in this study for predicting malaria cases.

2.3Disease forecasting using time series model:

Kim et al. [10] was to forecast the number of pneumonia cases in the province of Chiang Mai. An autoregressive integrated moving average (ARIMA) was used in data fitting to predict future pneumonia cases monthly. In Chiang Mai, 67,583 instances of Pneumonia were reported overall between 2003 and 2014, with a consistent monthly trend. The winter and rainy seasons in Thailand, which are February and September, had an increase in monthly pneumonia cases, whereas April through July saw a reduction (the period of the summer season to the early rainy season).

Kim et al. [10] used national health surveillance data;six meteorological parameters, including monthly rainfall and monthly average daily maximum temperatures, were used to determine the incidence of Pneumonia in children under the

age of five. A generalized linear model was used to calculate the magnitude of the influence of local and regional climate factors. Variation in the association between climate variables and Pneumonia is thought to represent the distribution of pneumonia risk factors and exposure in the neighbourhood.

Gao et al. [11] used Typhoid and paratyphoid incidence data from 2004 to 2016 in China were descriptively evaluated to investigate epidemiological aspects such as age-specificity and geographic dispersion. Compared to 2004, the cumulative incidence of both fevers decreased significantly in the country, by 86.6% for paratyphoid and 73.9% for Typhoid, respectively, in 2016. All age categories saw a decrease in cumulative incidence, but children aged 0 to 4 were the most susceptible in recent years. The grey model GM (1, 1) and seasonal autoregressive integrated moving average (SARIMA) model were used to extract the long-term patterns of the diseases. In this, SARIMA has the highest R^2 value and the lowest error value compared to others.

3.OBJECTIVES:

This work aims to predict and forecast the highest occurrence of infectious diseases from electronic patient health records. It is beneficial to health analysts to take prevention measures earlier.

- To predict the highest occurrence of infectious diseases based on the electronic patient health records using machine learning algorithms
- To find the deadly symptoms of infectious diseases depends on the patient's age using machine learning techniques.
- To forecast the future occurrence (year and day wise) of infectious diseases using deep learning algorithms in a big data environment.

4. RESEARCH GAP IDENTIFICATION:

With the increase in various infectious diseases, it was essential to find the disease prediction earlier and protect people from them. Many researchers only focus on finding the disease based on the symptoms. Otherwise, they concentrate on single disease prediction rather than multiple. In addition to that, the first time this study focuses on multiple disease prediction rather than single. Moreover, this study identifies the risk factors of patients based on their age.

5. ORGANIZATION OF THE THESIS:

Chapter 1:

Chapter 1 explains the introduction parts of infectious diseases and their types and discusses the essential components of contagious diseases. In addition to that, it examines how the spread of infectious diseases is related to climate factors. Moreover, it explains the transmission, incubation period and symptoms of each disease. It also discusses how machine learning, deep learning, and big data analytics are effectively used in healthcare sectors.

Chapter 2:

Chapter 2 illustrates the literature review related to the feature selection algorithms of disease prediction, machine learning algorithms of disease prediction and deep learning algorithms of disease forecasting. Several studies used optimization algorithms, Boruta and Xgboost, for feature selection. Random forest, XGBoost, and Decision tree classification algorithms are used for disease prediction. Most of the studies use LSTM-based time series forecasting algorithms for disease forecasting. These algorithms give the highest accuracy and low computational time.

Chapter 3:

Chapter 3 discusses the prediction of higher occurrence of infectious diseases. This study proposes a prediction model based on machine learning techniques for predicting seasonality diseases. In this proposed model, diseases like dengue, Malaria, Pneumonia, Typhoid, kala-azar, Japanese encephalitis, measles, and normal fever and cold infections are classified and predicted based on real-time data. This real-time data was collected from different regions around Madurai district from 2015 to 2020, with 63 features. The proposed model is a hybrid machine learning approach, which includes a feature selection and classification model for disease prediction. The Antlion Optimization Algorithm (ALO) is used for features selection, and for classifying the data with selected features, the Random Forest (RF) integrated with XG-Boost technique is used. Finally, based on the results of the data predictions, the performances are compared with the original data for validation. The proposed model is evaluated with the performance metrics like accuracy, precision, recall, specificity, and f1-score.

As per the experimental analysis, the proposed model has outperformed all the other compared machine learning algorithms. Figure 1 shows that the overall workflow predicts the highest occurrence of infectious diseases.

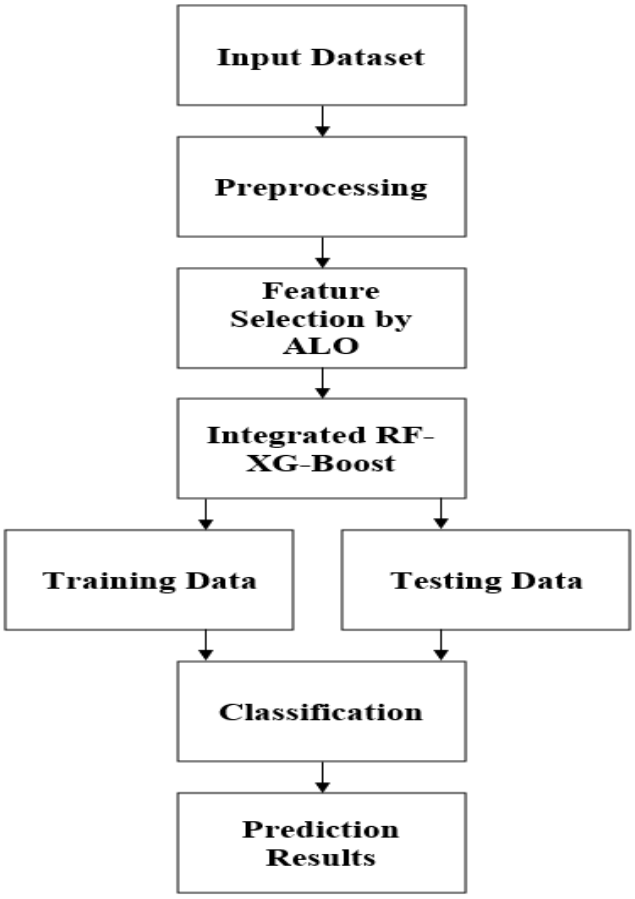

Figure 1: Overall workflow of disease occurrence prediction

The proposed model is a hybrid machine learning approach, which includes feature selection and classification techniques for disease prediction. The dataset contains 63 features, and the Antlion Optimization Algorithm (ALO) is used for feature selection. The Random Forest (RF) integrated with XG-Boost technique is used to classify the data with selected features. Finally, based on the results, the performance was compared with the original data for validation. The results of disease occurrences are shown in the figures.

A disease like Japanese encephalitis was predicted by the proposed model. The difference between the actual and predicted data was compared, as shown in Figure 2. As shown in the figure, the proposed model has close performance related to the actual data.

The proposed model successfully predicted the occurrence of Measles disease by comparing its predictions with the actual data in Figure 3. The results indicate that the model closely approximates the actual data, demonstrating its strong performance.

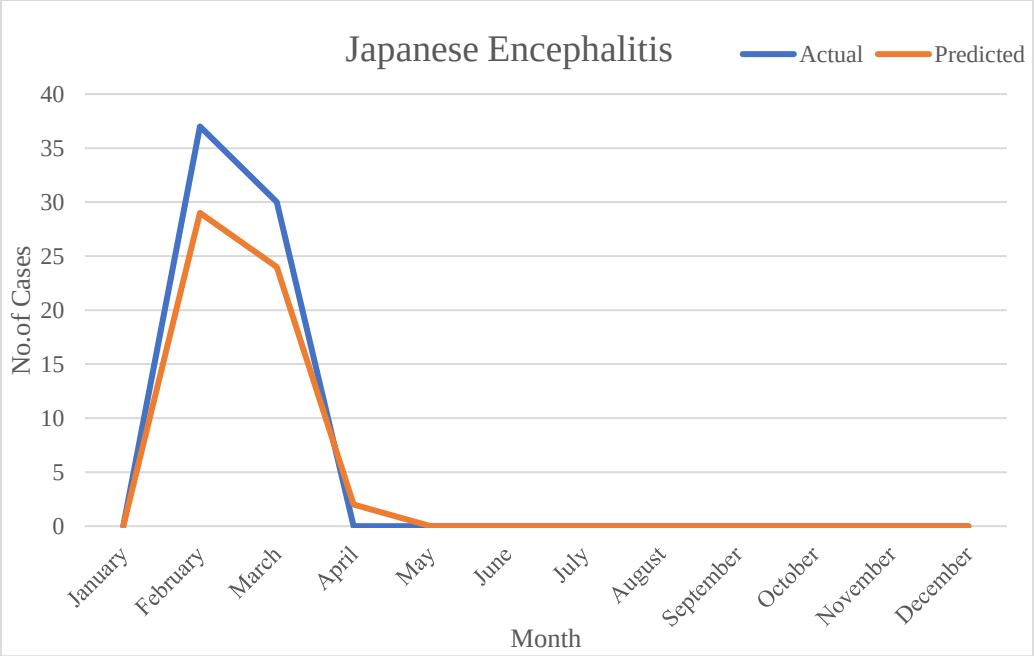

Figure 2: Proposed Model's Predicted Results for Japanese Encephalitis

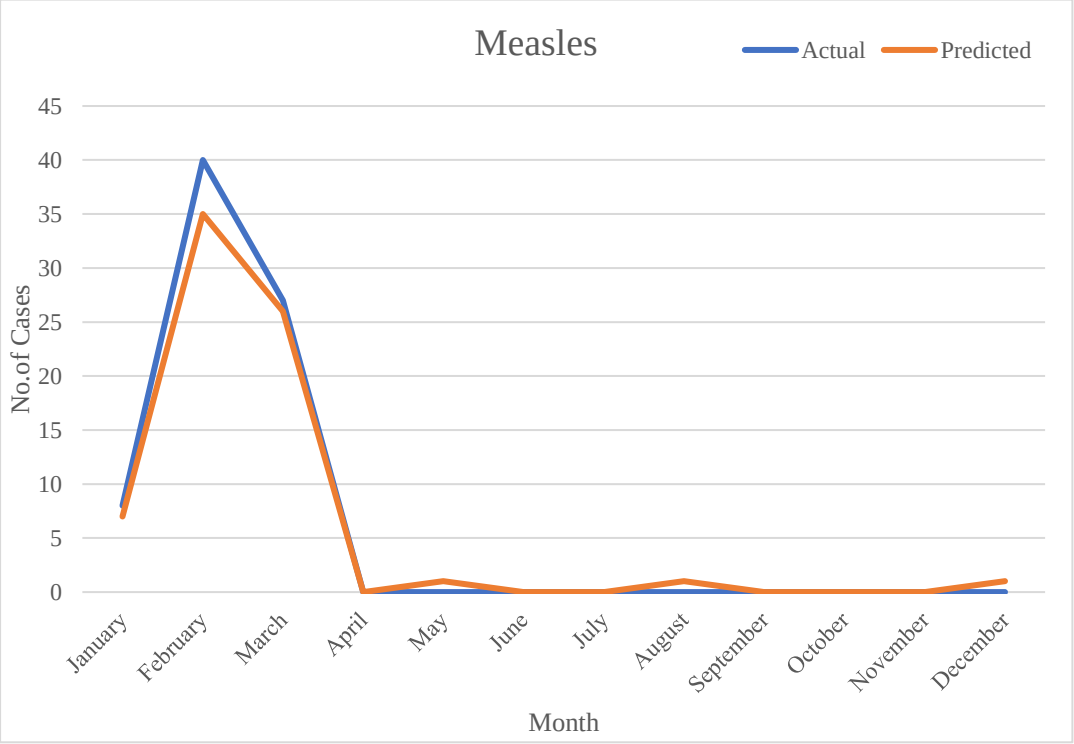

Figure 3: Proposed Model's Predicted Results for Measles

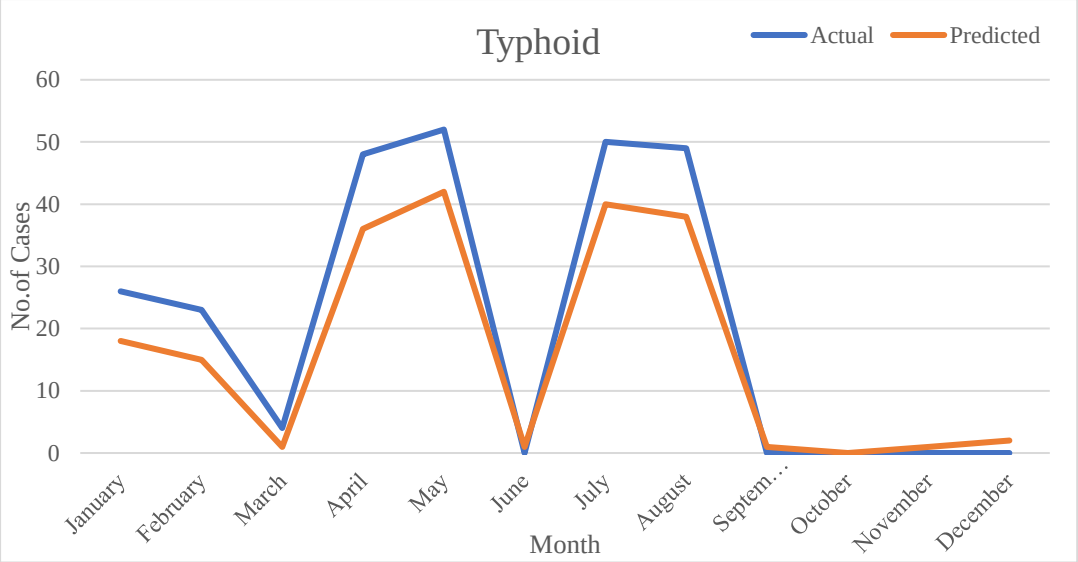

Figure 4: Proposed Model's Predicted Results for Typhoid

According to Figure 4, the proposed model accurately predicted the occurrence of Typhoid disease using the test dataset. A comparison between the actual and predicted data revealed minimal differences, indicating that the proposed model closely approximates the actual data in its predictions.

The predictions of the proposed model using the test dataset for diseases such as Pneumonia and Dengue are depicted in Figures 5 and 6. By comparing the actual and predicted data, it is evident from the figures that the proposed model exhibits a close performance to the actual data in both these predictions.

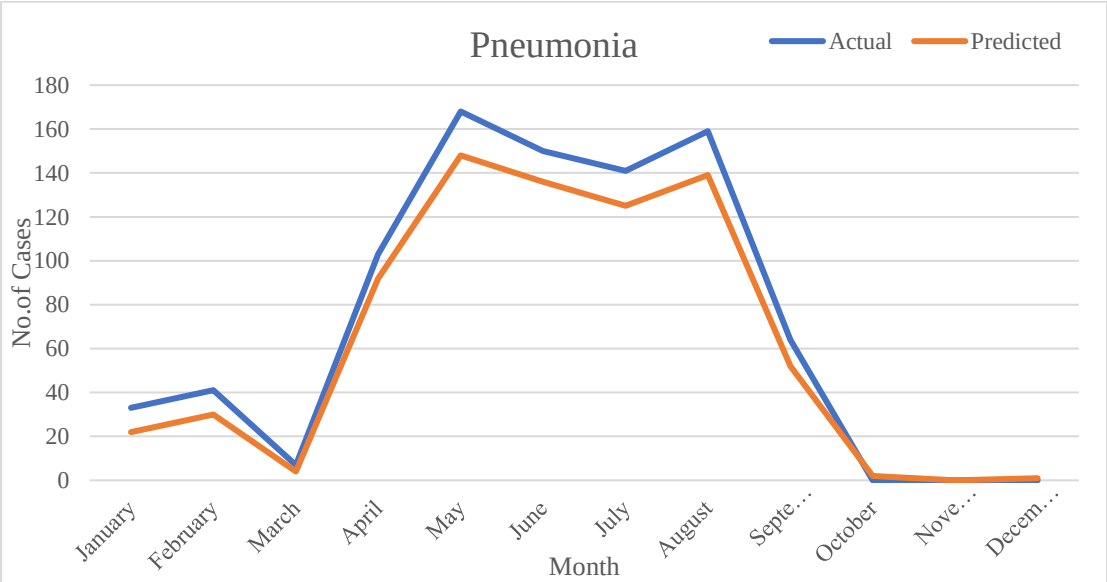

Figure 5: Proposed Model's Predicted Results for Pneumonia

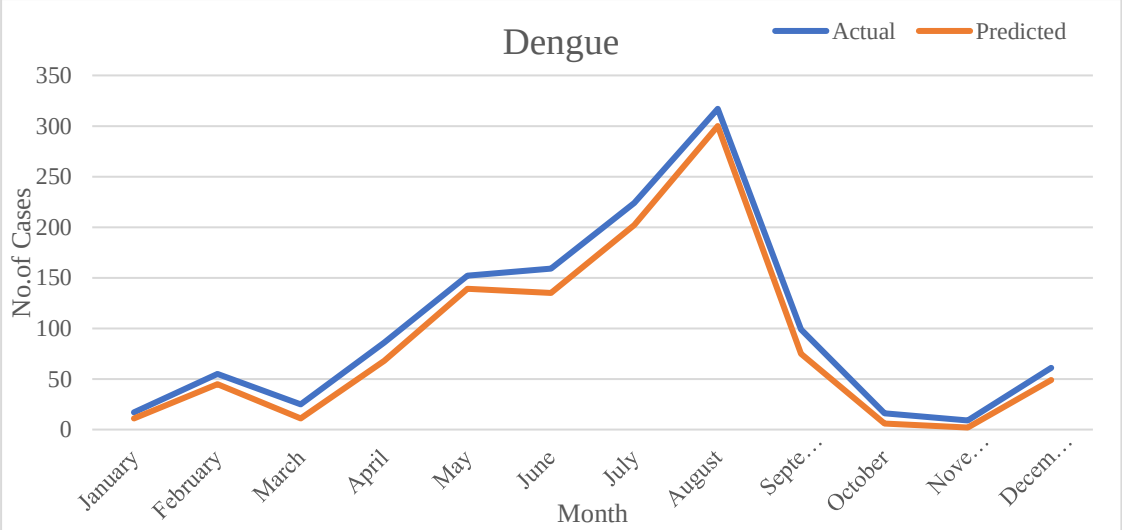

Figure 6: Proposed Model's Predicted Results for Dengue

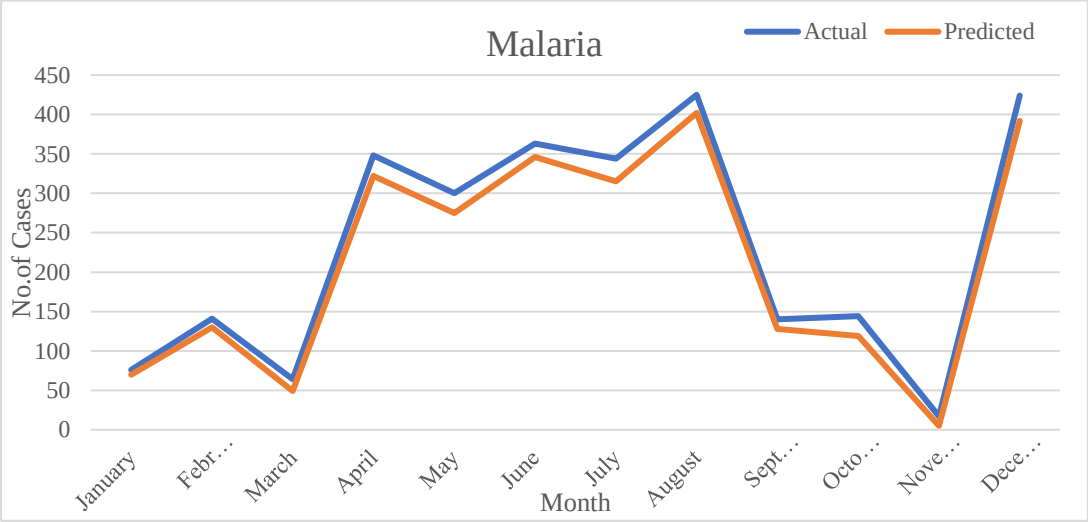

Figure 7: Proposed Model's Predicted Results for Malaria

In this research, Malaria is the most infected disease that occurs throughout the year annually. The proposed model has predicted the malaria disease from the test dataset with close performance to the actual data, as shown in Figure 7.Based on the results, find the month-wise disease occurrence and seasonal patterns of various diseases. Based on the results, this study has identified the top four most ailments which occurrences with the highest count.

Chapter 4:

Chapter 4 identifies the deadly symptoms of four infectious diseases found in Chapter 3. Machine learning is the most popular technique for predicting the early signs of ailments in the medical industry. Dengue, Malaria, Pneumonia and Typhoid cause

severe consequences among people. In particular, Dengue and Malaria generate severe effects compared to the other two infections. These diseases affect people of different ages. So, early prediction of risk signs and affected age patterns is very beneficial for health analysts to prevent patients from risk. This research discusses the early prediction of age-wise symptoms of four infectious diseases such as Dengue, Malaria, Pneumonia and Typhoid. This study proposes a novel feature selection method Patient Risk Factor – Boruta XGBoost (PRF-BXGBOOST) and a classification method Chi Square Automatic Interaction Detection Tree (CT) for identifying the patient's age-wise symptoms. The performance evaluation of this new novel approach is determined by using the following parameters accuracy, precision, recall and F1-score. The proposed model outperforms the other machine learning models based on the performance comparison. Figure 8 shows the overall workflow of Age-wise disease symptom prediction.

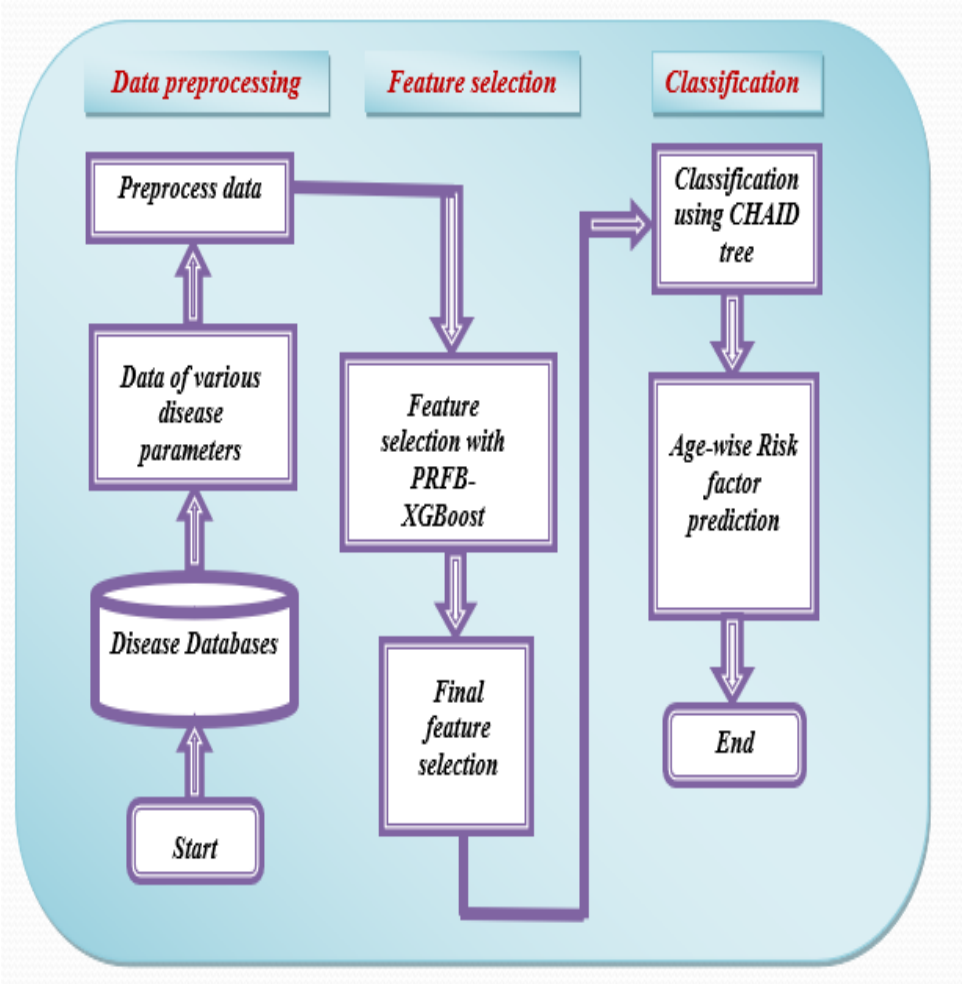

Figure 8: Overall workflow of Age-wise disease symptom prediction

This result analyses the age-wise deadly symptoms of various infectious diseases like Dengue, Malaria, Pneumonia and Typhoid. The age category is divided into three classes, such as child (0 to 15), young (15 to 50) and old (50 to 89).

Figure 9 shows the age-wise symptoms of dengue patients. Dengue fever has several normal symptoms. But this study identifies the age-wise dangerous symptoms. From Figure 9it can be understood that nose-bleeds, blood vomiting and muscle pain are the most dangerous symptoms for children. These are the deadly symptoms of children. Most of the children were infected by dengue disease compared to the other two age groups (young and old). Besides, body pain and blood vomiting are dangerous symptoms of old and young people. Older people have other diseases like blood pressure, sugar etc. It leads to the organ failure stage. If older people have organ failure, the patient's risk is very high.

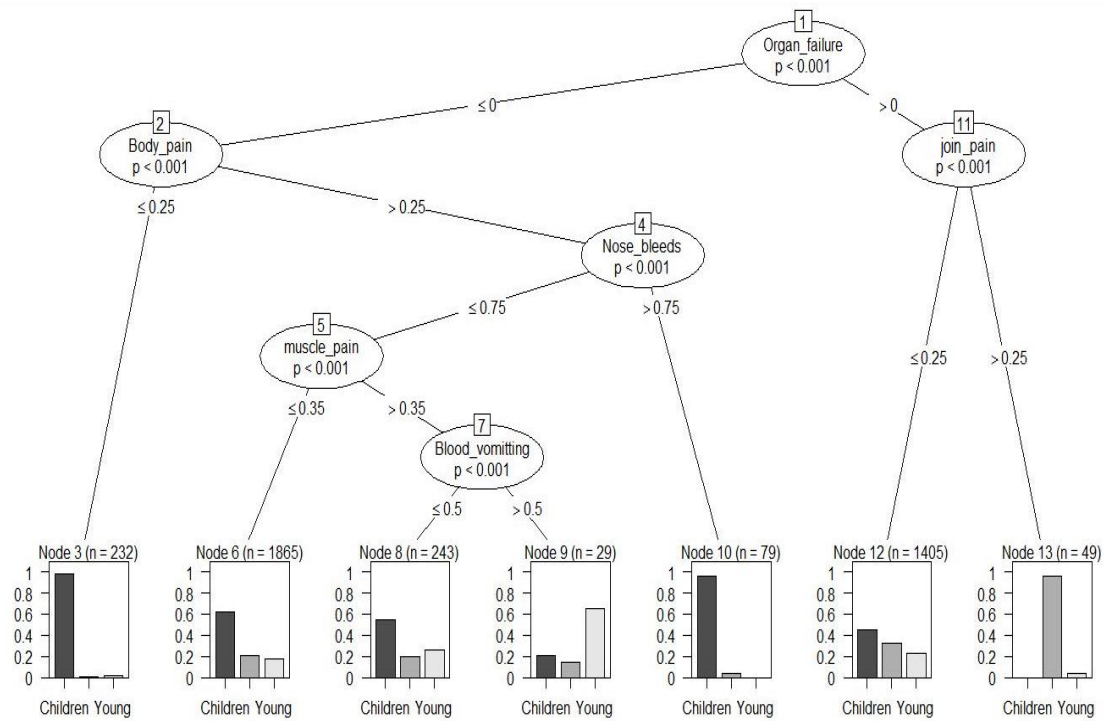

Figure 9: Age-wise risk symptoms of Dengue with CHAID

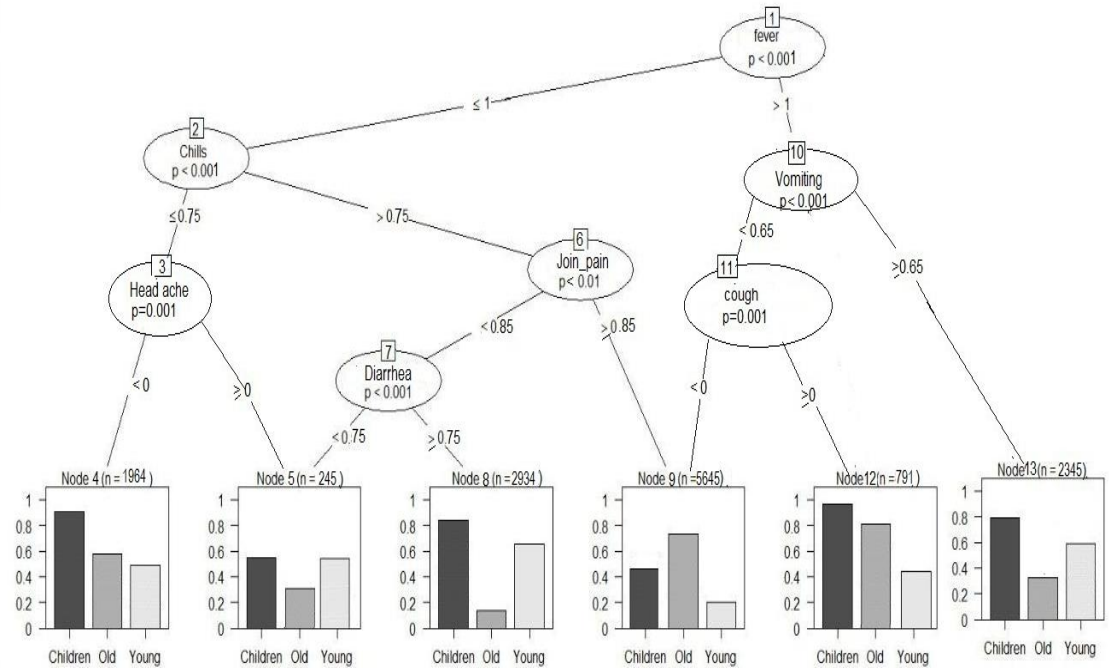

Figure 10: Age-wise risk symptoms of Malaria with CHAID

Figure 10 depicts the age-wise deadly symptoms of Malaria. From the figure, it is understood that fever, chills, diarrhoea, cough and vomiting are the major symptoms of children. Fever, cough, vomiting and joint pain are the deadly symptoms of older people. Finally, fever, chills, headache and diarrhoea are the most dangerous symptoms of young people. Most of the children were infected by Malaria when compared to the other two age groups.

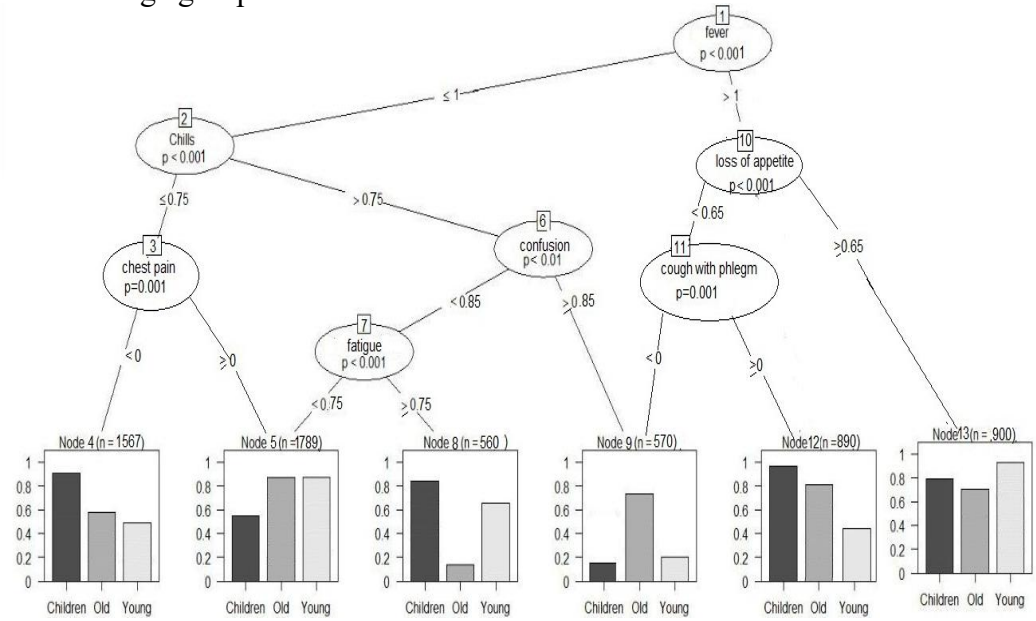

Figure 11: Age-wise symptoms of Pneumonia with CHAID

Figure 11 exposes the age-wise risk symptoms of Pneumonia. Based on the above figure, it is observed that fever, cough with phlegm, loss of appetite, fatigue and chills are deadly symptoms in children. Fever, chest pain, confusion and chills are the dangerous symptoms of older people. Older people only have confusion symptoms. Fever, loss of appetite, cough with phlegm, fatigue and chest pain are the major symptoms of young people. Most of the symptoms of Pneumonia are common to all age groups. But each group has unique symptoms depending on the age.

Figure 12 presents the age-wise deadly symptoms of typhoid patients. The main symptoms in kids and teenagers include fever, stomach ache, vomiting and diarrhoea. Only a smaller number of older people were infected by typhoid disease. Many of the children were affected by the Typhoid.

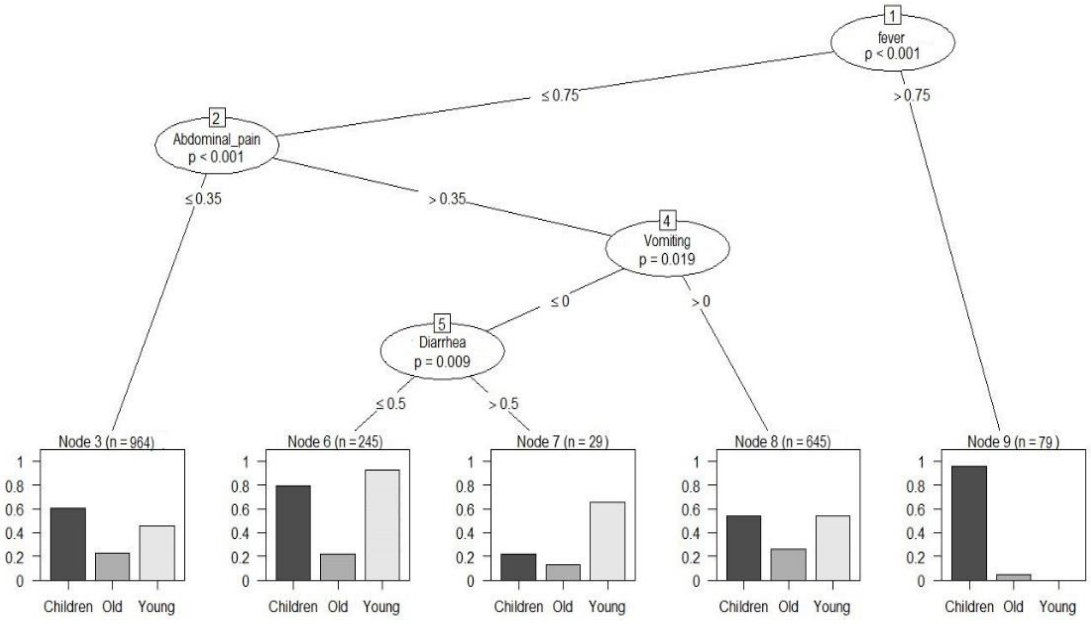

Figure 12: Age-wise symptoms of Typhoid with CHAID

Chapter 5:

Chapter 5 forecasts the future occurrence (day-wise and year-wise) of four infectious diseases. The main aim of this study is to forecast the occurrence of the next 15 years of seasonal ailments like Dengue, Malaria, Pneumonia and Typhoid by combining patient records and climatic parameters. This study considers three climatic elements like minimum temperature, maximum temperature and rainfall. In this study, six machine learning algorithms have been used such as LSTM (Long Short-Term

Memory), Bi-directional LSTM, GRU (Gated Recurrent Unit), RNN (Recurrent Neural Network), ARIMA (Auto Regressive Integrated Moving Average) and Deep Stacked Bi-directional LSTM. The performance metrics Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are used to compare the performance of the proposed model. The proposed model, Deep Stacked Bi-directional LSTM, gives the best forecasting results compared to other models. The study's findings suggest a trustworthy early detection system that could predict the appearance of these infectious diseases. It will help the relevant authorities carry out exact preventive measures. Figure 13 represents the overall process of disease forecasting.

This result focused on two types of disease forecasting. The two types are day-wise forecasting and year-wise forecasting. The proposed method forecasts disease occurrence based on the previous occurrence of epidemics and climate factors such as maximum temperature, minimum temperature and rainfall.

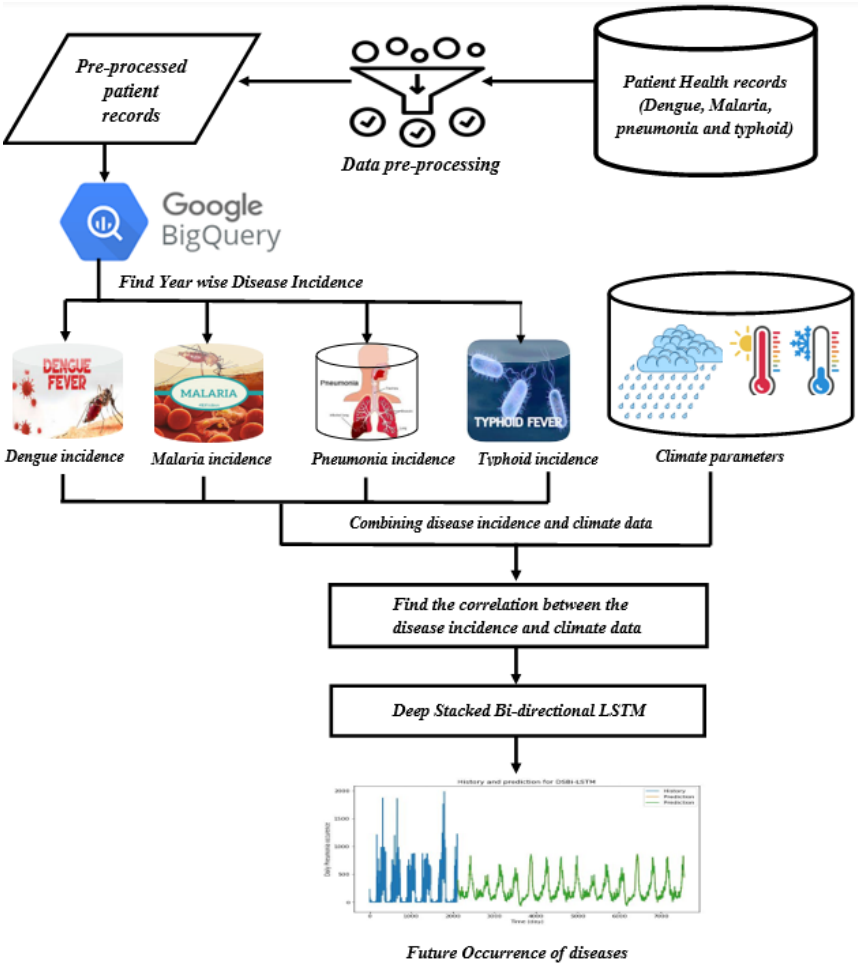

Figure 13: **Overall work process of disease forecasting**

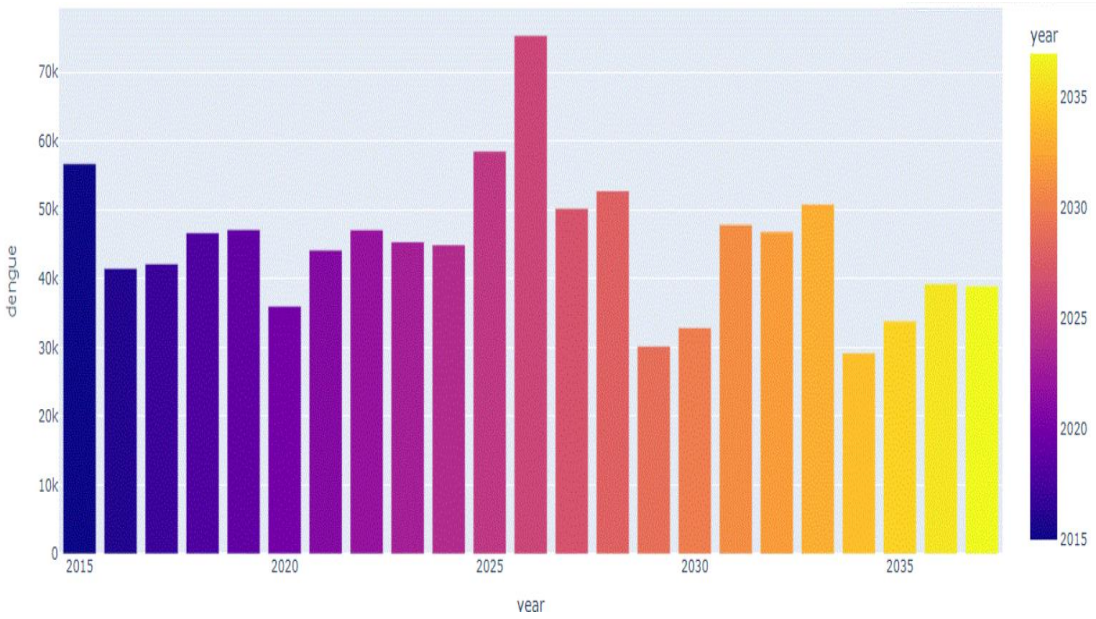

Figure 14: Year-wise future forecasting of dengue disease

Figure 14 depicts the next 15 years of Dengue occurrence in Madurai district from 2022 to 2037. Based on the figure, Dengue occurrence will reach its peak in the year 2026, with the occurrence of Dengue cases going up to more than 70,000, and it will reach the least in the year 2034.

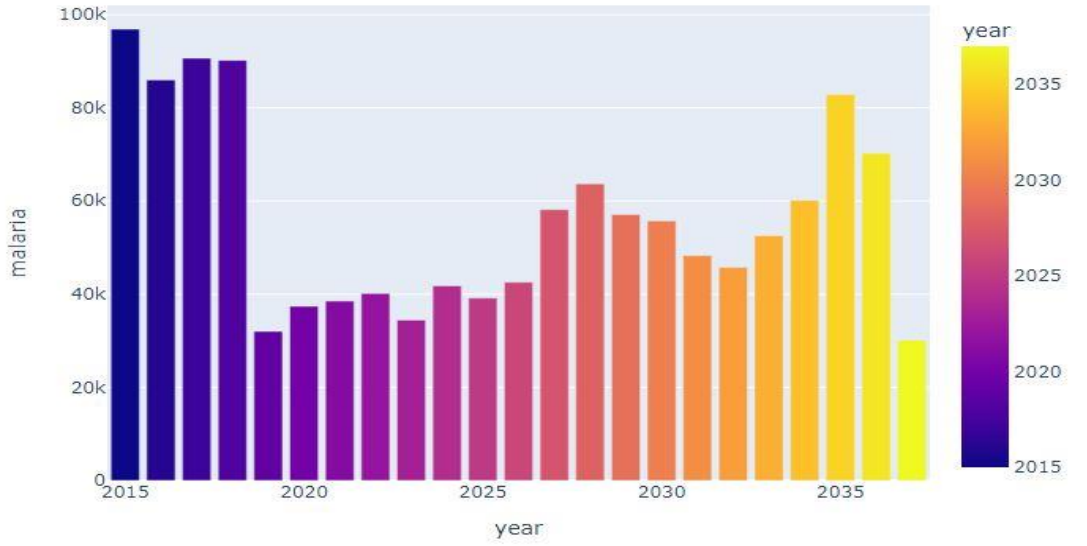

Figure 15:Year-wise future forecasting of Malaria disease

Based on the data presented in Figure 15, the proposed model provides year-wise predictions for the next 15 years. The figure indicates that the occurrence of Malaria is projected to reach its highest point in the year 2035, with the number of

malaria cases exceeding 80,000. Conversely, the figure also reveals that the occurrence of Malaria will reach its lowest point in the year 2037.

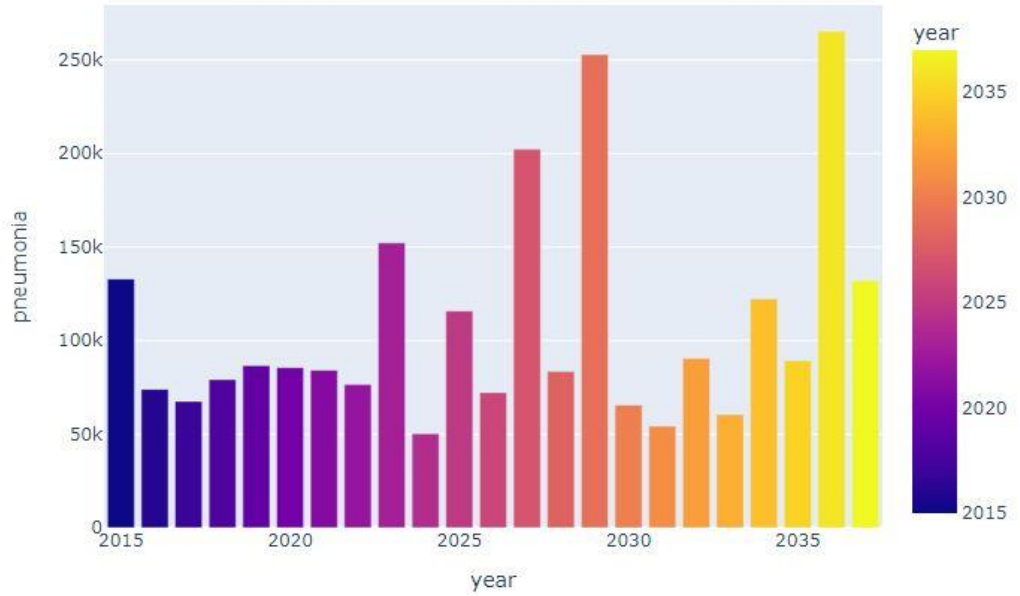

Figure 16: Year-wise future forecasting of pneumonia disease

Figure 16 illustrates the projected occurrence of Pneumonia in Madurai district for the period from 2022 to 2037. According to the figure, the incidence of Pneumonia is expected to reach its highest point in 2036, with the number of pneumonia cases surpassing 250,000. Conversely, the figure also indicates that the occurrence of Pneumonia will be at its lowest in 2024.

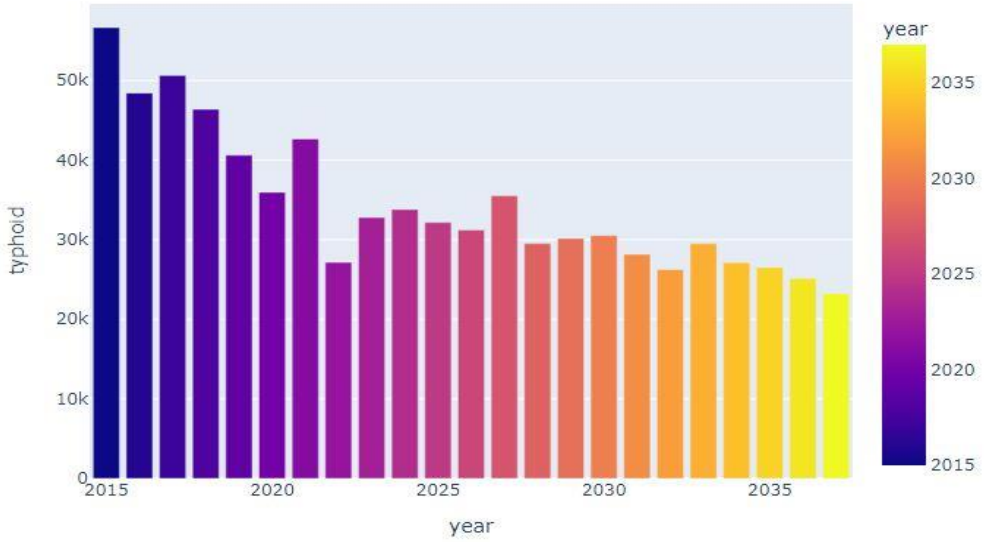

Figure 17: Year-wise future forecasting of typhoid disease

Figure 17 displays the predicted year-wise occurrences of Typhoid using the proposed model for the next 15 years. Based on the figure, it can be observed that the incidence of Typhoid is projected to reach its highest point in 2027, with the number of typhoid cases exceeding 35,000. Furthermore, the figure indicates that the occurrence of Typhoid will be at its lowest in 2037.

Chapter 6:

Chapter 6 explores the conclusion of the present work. In this research, a prediction model based on machine learning techniques was proposed for predicting seasonality diseases. In this proposed model, diseases like Dengue, Malaria, Pneumonia, Typhoid, Kala-azar, Japanese encephalitis, measles and normal fever and cold infections were classified and predicted based on real-time data. Based on the Performance evaluation, the proposed model (Integrated Random Forest – XGBoost) has outperformed all the other compared models in terms of disease prediction. During the initial phase, an analysis revealed the highest incidence of four infectious diseases: Dengue, Malaria, Pneumonia and Typhoid. The second phase identified the risk factors of these four diseases. In that, the PRFBXGBOOST-CHAID technique performs well. It accurately classifies the symptoms of patients in Madurai district based on their age group. This work will be very helpful to the doctors in identifying the patients' risk level depending on the age group. Early age-wise risk prediction is beneficial to rescue patients from death. In the third phase, forecast the next 15 years of disease occurrence based on climatic factors such as temperature and rainfall. Based on the result analysis, Pneumonia will occur in the highest count in the Madurai district during the year 2022-2037 when compared to the other three diseases, Dengue, Malaria and Typhoid. This research benefits medical researchers and public health departments to take preventive measures for the forthcoming years.

6. References:

[1] E. Smith, MSN and B.C. WHNP, The effect of potential climate change on infectious disease presentation, *The journal of nurse practitioners*, **15** (6) (2019), 405 -409.

[2] X.X. Wu, H. Tian, L.D. Gao, H.N. Liu and B. Xu, Impact of global change on transmission of human infectious diseases, *Sci.China Earth Sci*, **57** (2) (2020), 189 – 203.

[3] Sharma et al, A heterogeneous ensemble forecasting model for disease predictions. *New Gen Comput*, **39** (3-4) (2021), 1-5.

[4] S. Grampurohit and C. Sagarnal, Disease predictions using machine learning algorithm, *International Conferences for Emerging Technology (INCET)*, **4** (3) (2022), 1-7.

[5] P. Dutta, S. Paul, A.J. Obaid, S. Pal and K. Mukhopadhyay, Feature Selections based Artificial Intelligence Technique for the Predictions of COVID like Diseases, *Journal of Physics: Conference Series*, **1963** (1) (2021), 1-11.

[6] F.E. Ayo, J.B. Awotunde, R.O. Ogundokun, S.O. Folorunso and A.O. Adekunle, A decision support system for multi-target disease diagnosis: A bioinformatics approach. *Heliyon, cell press*, **6** (2020), 1-14.

[7] H.H. Thary and K. Azidan, A framework Questionnaire for diagnosing infectious disease using machine learning techniques INTCSET 2020, *IOP conference series: Materials science and Engineering, IOP Publishing*, **1094** (2020), 1-10.

[8] M. Mariki, E. Mkoba and N. Mduma, Combining clinical symptoms and patient features for malaria diagnosis: machine learning approach, *Applied Artificial Intelligence, Taylor & Francis*, **36** (1) (2022), 1-25.

[9] Y.A. Adamu and J. Singh, Malaria prediction model using machine learning algorithms. *Turkish journal of computer and mathematics Education*, **12** (10) (2021), 7488-7496.

[10] Kim et al, Effect of Climate Factors on the Childhood Pneumonia in Papua New Guinea: A Time-Series Analysis. *International journal of Environmental Research and Public Health. MDPI*, **13** (2) (2021), 213.

[11] J. Gao, J. Li and M. Wang, Time series analysis of cumulative incidences of typhoid and paratyphoid fevers in China using both Grey and SARIMA models, *PLOS ONE*, **15** (10) *(*2020), 1-14.

7. List of Publications:

[1] K. Indhumathi and K. Sathesh Kumar, A review on prediction of seasonal diseases based on climate change using big data, *Materials Today Proceedings& Elsevier*, **37** (2021), 2648-2652.

[2] K. Indhumathi and K. Sathesh Kumar, Seasonal Infectious Disease Prediction based on Electronic Patient Health Records using Boosted Random Forest Algorithms, *2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, Greater Noida, India, (2022), 2025-2032,

[3] K. Indhumathi and K. Sathesh Kumar, Efficient Prediction of Seasonal Infectious Diseases Prediction using Hybrid Machine Learning Algorithms and Feature Selection Techniques, *International Journal on Artificial Intelligence Tools, World Scientific*, (2023).

[4] K. Indhumathi and K. Sathesh Kumar, Prediction of Seasonal Infectious diseases based on hybrid machine learning approaches, *Multimedia Tools and Applications, Springer*, (2023).

[5] K. Indhumathi and K. S. Kumar, Prediction of Infectious Diseases based on Age and Gender using OCSVM-CDT, *Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India*, (2023), 1269-1276.